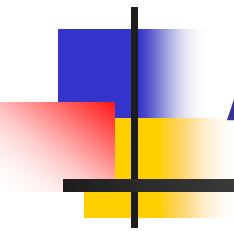
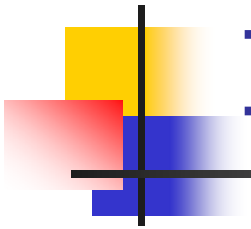


ABS : Adaptive Borders Search

Frédéric Flouvat[1]

Travail réalisé avec
Fabien De Marchi[2] et Jean-Marc Petit[1]

[1] LIMOS - Université Blaise Pascal – Clermont-Ferrand, France
[2] LIRIS – Université Claude Bernard – Lyon, France



Introduction

- Travail en cours
- Deuxième année de thèse
- Objectif : Découverte des motifs fréquents maximaux d'une base de données de transactions
  - Généralisable à d'autres problèmes de fouille de données [MANNILA ET TOIVONEN 97]
    - représentable par des ensembles
    - prédicat anti-monotone

Les motifs fréquents maximaux par rapport à l'inclusion

- Permet de représenter l'ensemble des fréquents *FI*

- Bordure positive $Bd^{+}(FI)$
  - Ensemble des motifs fréquents de FI maximaux (MFI) par rapport à l'inclusion

- Bordure négative $Bd^{-}(FI)$
  - Ensemble des motifs non fréquents minimaux par rapport à l'inclusion

Problématique

- Performance des algorithmes de découverte des MFI
  - workshop **F**requent **I**temset **M**ining **I**mplementations 2003 (ICDM'03):
    - Apriori reste compétitif pour découvrir les MFI lorsque la taille des MFI reste « petite »

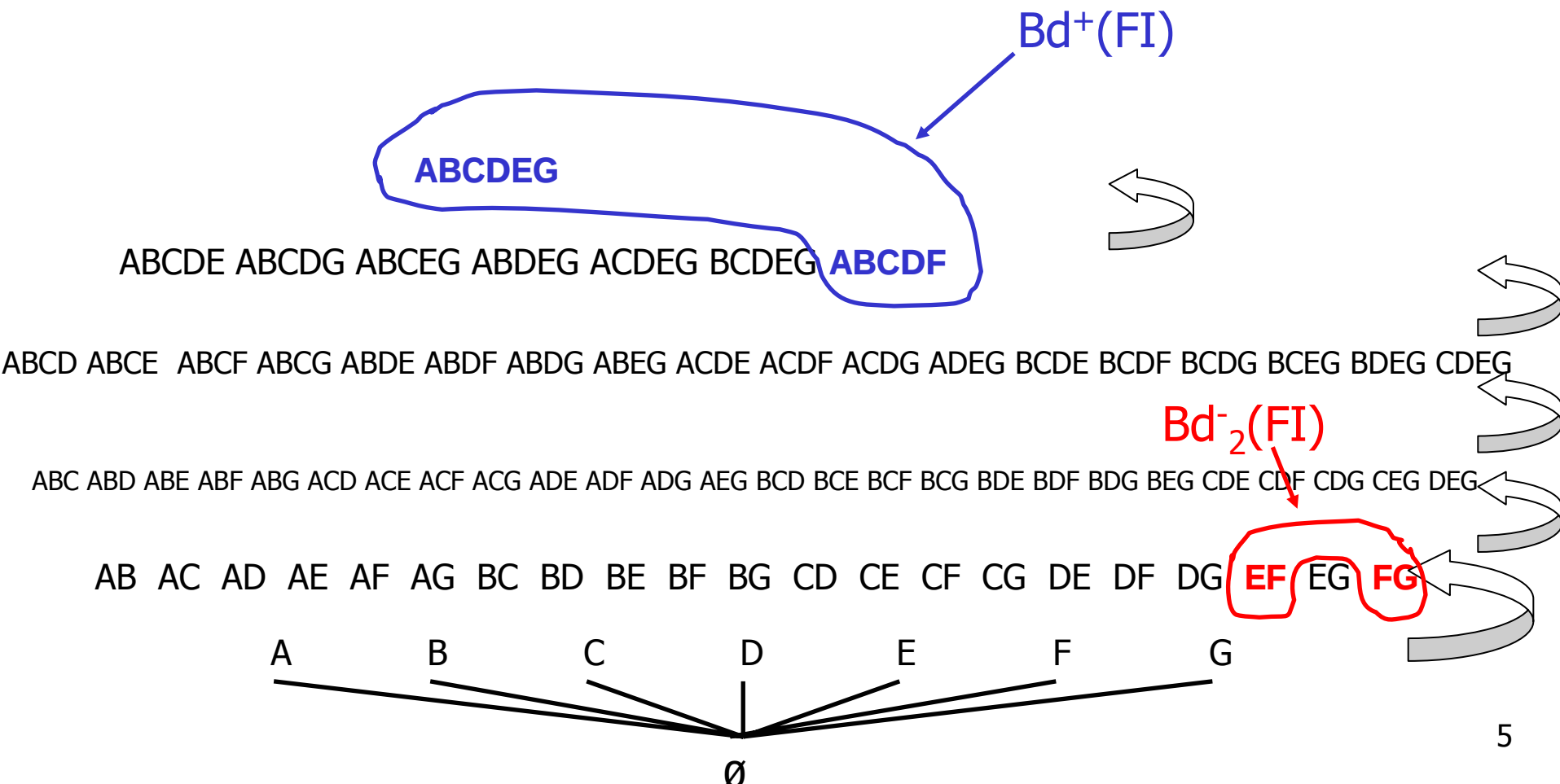
- Problème quand de grands MFI existent

- Idée d'ABS:

  « lorsque Apriori rencontre des difficultés, on change de stratégie »

- Deux principaux problèmes :
  - Comment décidé dynamiquement « lorsque Apriori rencontre des difficultés »
  - Quelle stratégie ?

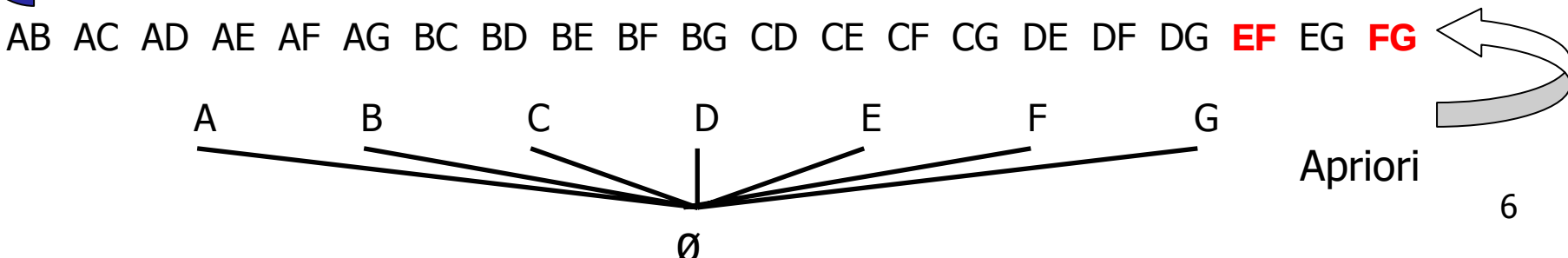
Exemple de découverte des MFI avec Apriori



Exemple utilisant les informations déjà découvertes

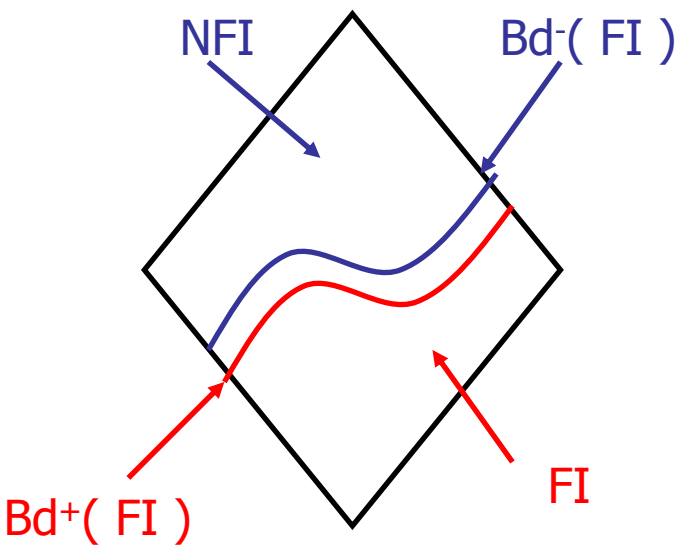
=> 1 itération au lieu de 4 iterations avec Apriori

Recherche des plus grands motifs ne comprenant pas {EF, FG}



Résultats théoriques sur les bordures (1)

- Liens entre les bordures
  - Transversaux minimaux d'un hypergraphe [MANNILA ET TOIVONEN 97]



$$FI = \{ X \subseteq R \mid X \text{ est fréquent} \}$$
$$NFI = \{ X \subseteq R \mid X \text{ est non fréquent} \}$$

$$Bd^-(FI) = TrMin(\overline{Bd^+(FI)})$$

Résultats théoriques sur les bordures (2)

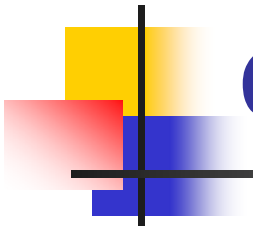
- $TrMin(TrMin(H)) = H$ [BERGE 74]

  avec H un hypergraphe

- Caractérisation de la bordure positive :

$$Bd^{+}(F I) = TrMin(Bd^{-}(F I))$$

Stratégie fondée sur des dualisations



- Dualisation = calcul des transversaux minimaux
- Dans le même esprit que l'algorithme Dualize and Advance [GUNOPOULOS D. et al. 03]
- Stratégie choisie à ce jour:
  - alterne des dualisations entre les deux bordures

Intérêts de ABS

- Caractérisation exacte de la bordure positive optimiste
- Approche non fondée sur une heuristique comme MaxMiner, Mafia, GenMax
  - Prix à payer : le coût de la dualisation
- Approche adaptative
- Remarque: très sensible à la première dualisation

Comportement adaptatif

- Quand Apriori doit-il s'arrêter ?
  - Comportement adaptatif s'appuyant sur les informations déjà découvertes
  - Principaux paramètres pour un niveau k
    - Taille de la bordure négative
      - $| Bd^-_k | < | F_k |$ ou $| Bd^-_k | \approx | F_k |$
        - coût de la dualisation *vs* coût de générer et compter avec Apriori
    - Ratio $| F_k | / | C_k |$
      - Si proche de 1, dualiser doit être intéressant
- L'algorithme peut ne faire aucune dualisation, i.e. ABS = Apriori

Caractérisation des bordures (1)

- Bordure positive optimiste Bd^{+}_{opt} :
  - $Bd^{+}_{opt} = TrMin(Bd^{-}_{k}(FI))$

    avec $Bd^{-}_{k}(FI)$ l'ensemble des motifs de la bordure négative découverts à l'itération k

  - $X \in Bd^{+}_{opt}$, X appartient au MFI ou X couvre un MFI

- Rq: plus le nombre de motifs de $Bd^{-}(FI)$ trouvés est important, plus on va tendre vers $Bd^{+}(FI)$

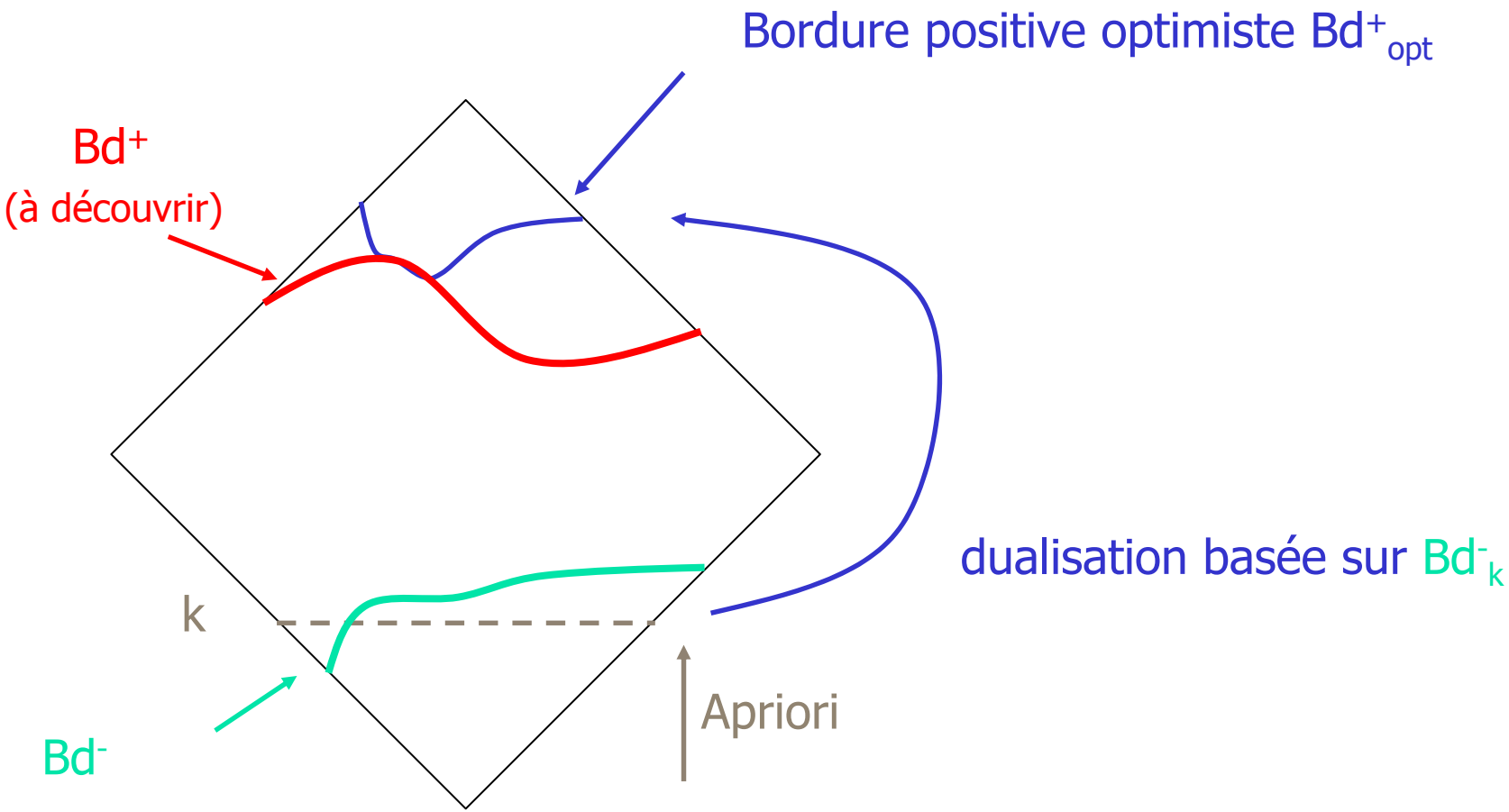
Caractérisation des bordures (2)

- Motifs candidats C_{i+1} :
  - $C_{i+1} = \overline{TrMin(Bd^+_i (FI))} - U_{j \leq i} C_j$

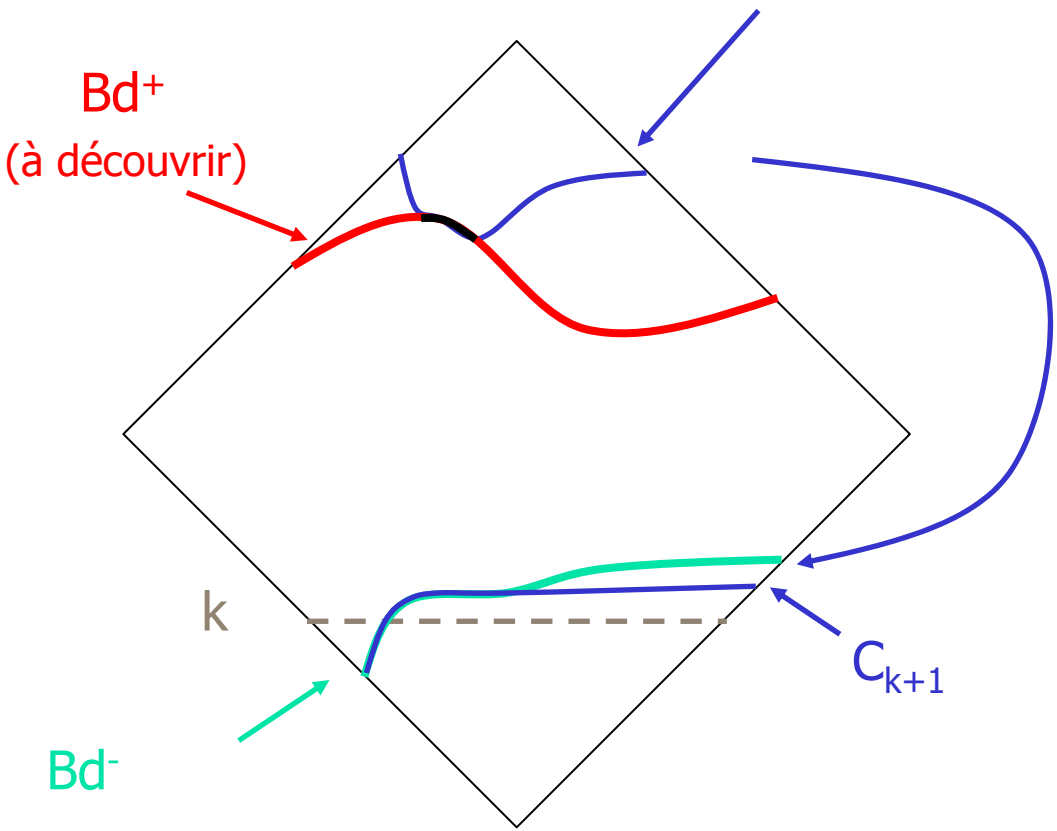
  avec $Bd^+_i (FI)$ l'ensemble des MFI découverts jusqu'à l'itération i

- Ensemble des plus petits motifs ne faisant pas partie d'une partie de l'espace de recherche déjà caractérisé

Une nouvelle stratégie

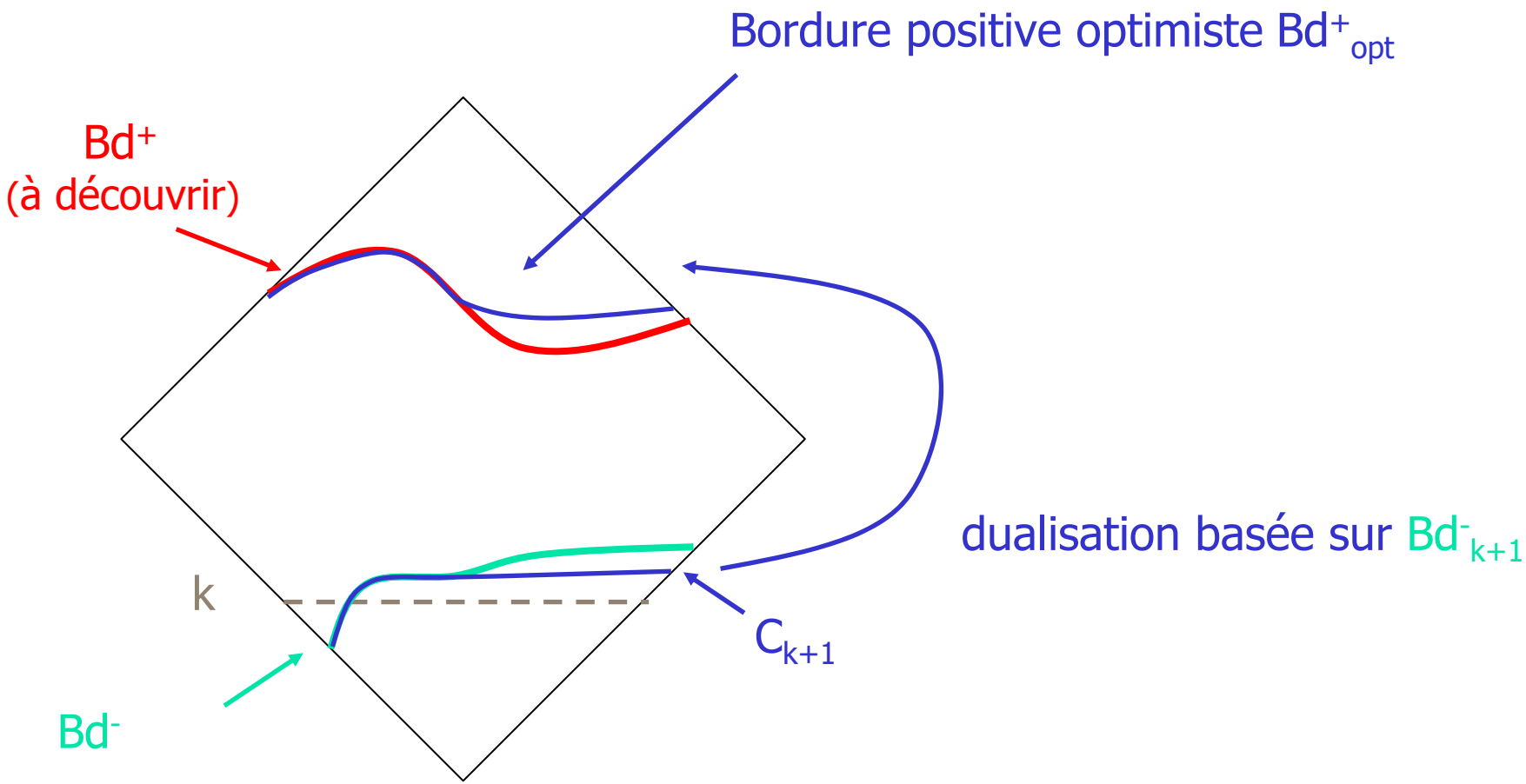


Une nouvelle stratégie

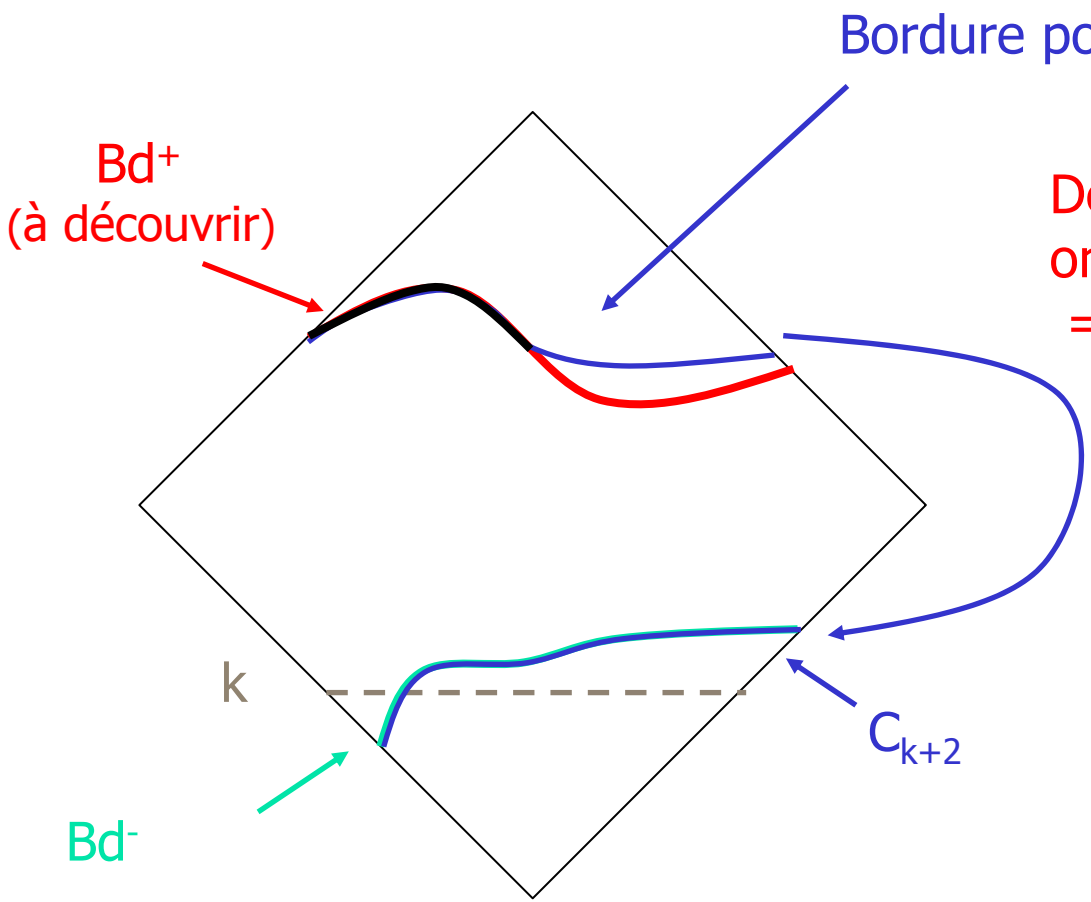


Des motifs non fréquents ont été trouvés
=> réitère dualisation basée sur les MFI trouvés

Une nouvelle stratégie



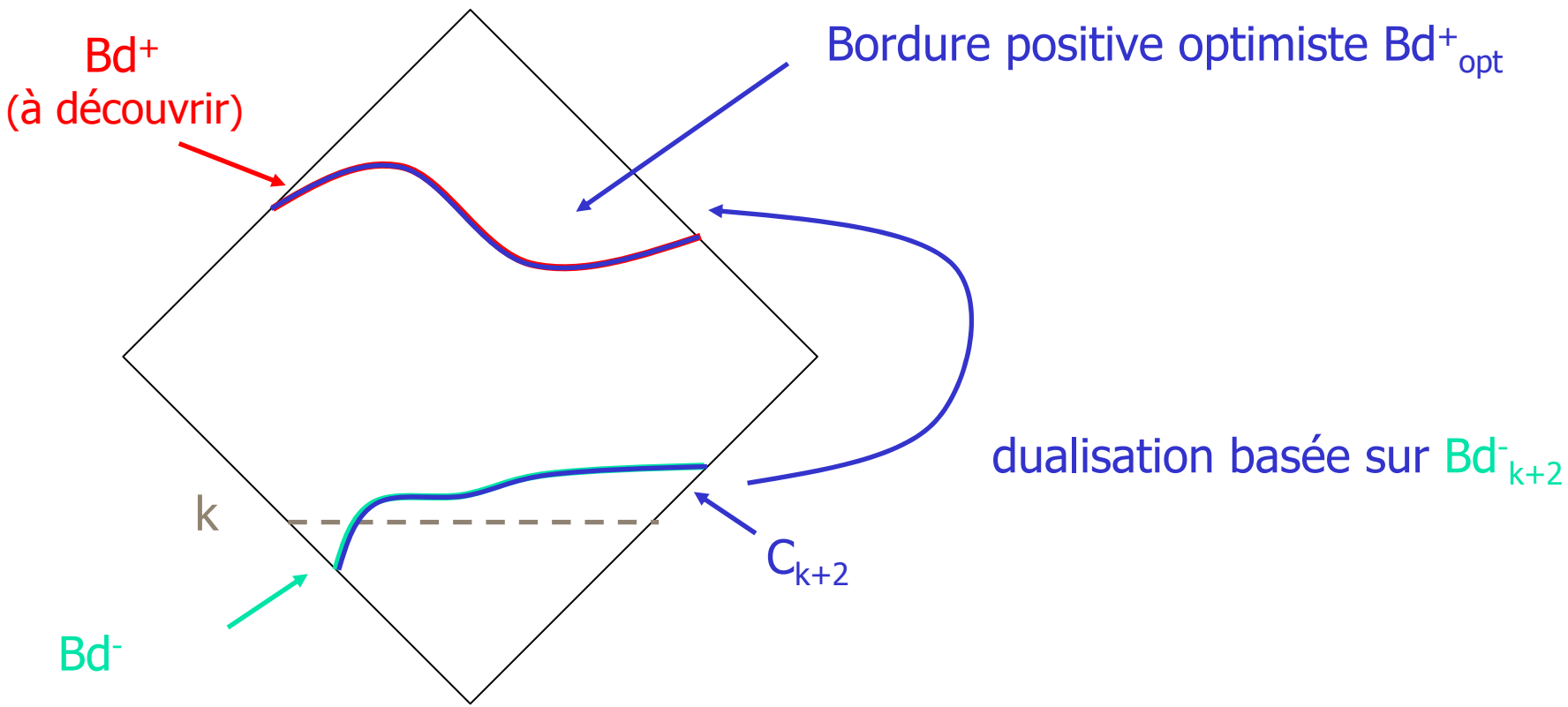
Une nouvelle stratégie



Des motifs non fréquents ont été trouvés
=> réitère

dualisation basée sur les MFI trouvés

Une nouvelle stratégie



Etat de l'art (1)

- MaxMiner [BAYARDO R.J. 98]
  - Parcours par niveaux
  - "sauts" :
    - test le support du plus grand motif apparaissant dans le sous-arbre du motif en cours
  - Calcul d'une borne inférieure pour le support des candidats

- Mafia [BURDICK D. et al. 01] et Genmax [GOUDA K. et ZAKI M. 01]
  - Parcours en profondeur
  - Stratégies effectuant des sauts approximatifs
  - Différentes optimisations pour le stockage de la base de données et pour limiter la taille de l'arbre

Etat de l'art (2)

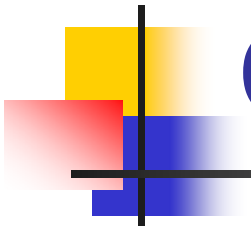
- Dualize and Advance [GUNOPOULOS D. et al. 03]
  - Dualisation à partir de motifs fréquents
  - À partir du premier motif fréquent trouvé, recherche du motif fréquent maximal le couvrant par un parcours en profondeur
  - Autant d'itérations que de motifs fréquents maximaux

Quelques résultats expérimentaux (1)

- Implémentation en C++ s'appuyant sur les implémentations d'Apriori de C. Borgelt et B. Goethals
- Utilisation des arbres préfixés (trie) pour représenter les ensembles de motifs
- Jeux de données réels et synthétiques (FIMI)
- Pas la meilleure implémentation (FIMI'04) 
  - Travail en cours
  - Comptage, dualisation… pas assez performants

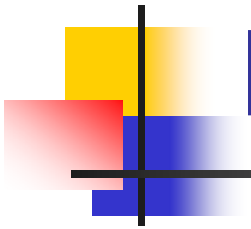
Quelques résultats expérimentaux (2)

- Mais a de meilleures performances que IBE (FIMI'03)
  - Algorithme dérivé de Dualize and Advance
  - Seule implémentation disponible utilisant le concept de dualisation
  - Le nombre de dualisation peut être amélioré
- ABS = Apriori sur le jeu de données Retail
  - Aucune dualisation n'est faite
- ABS peut améliorer Apriori par un facteur 10 sur Connect, Pumsb*



Conclusion

- Nouvelle approche adaptative qui combine la force des algorithmes
  - Apriori
  - Dualize and Advance
- Coût élevé de la dualisation pour de grands ensembles
  - Meilleure performance avec des heuristiques
- Approche générique pouvant être utilisé dans d'autres problèmes de fouille de données
  - représentable par des ensembles
  - prédicat anti-monotone
- Amélioration d'Apriori



Perspectives

- Améliorer les aspects adaptatifs
  - 1ère dualisation
  - Utiliser les MFI « presque vrais »
- Principal goulot d'étranglement pour les MFI :
  - Coût de la dualisation
- Tirer avantage des informations découvertes avant la première dualisation
  - E.g. règles d'association exactes, motifs fermés fréquents

Questions ?